Organization and Exploration Fined-grained Historical Knowledge on Contemporary China Based on Semantic Mining

ZHANG Zhixiong[1], WANG Ying[1], SUN Hui[2] & LEI Feng[2]

1 National Science Library, Chinese Academy of Sciences, Beijing 100190, China

2 The Institute of Contemporary China Studies, Beijing 100009, China

ABSTRACT

China has a huge volume of historical resources on its contemporary history. Lots of valuable knowledge are hidden in those resources and cannot be utilized easily. It is an urgent problem to mine the implicit semantic knowledge scattered in a large number of historical resources and to reorganize the historical knowledge and facts in a fined-grained manner, so that can help user to explore the historical knowledge for research and education.

This paper proposes a method, which is called "Mining down, Organizing up", to semantically represent and organize historical knowledge on contemporary China hidden in historical encyclopedia text. Based on the proposed historical ontology of contemporary Chinese, this method extracts knowledge objects and facts from the unstructured historical text items by utilizing text mining technologies, represents the historical knowledge in semantically enriched way, and interlinks the related historical knowledge objects and facts to form a historical knowledge network of the contemporary China. By mining the historical facts and the historical knowledge network, the authors get more valuable patterns from the historical knowledge which could be used to form the new organization scheme to reorganize the historical knowledge in a more vivid way.

Based on this method, the authors developed a system which can represent and organize historical knowledge of contemporary China in a fined-grained manner, support user to explore historical knowledge by providing functions such as semantic retrieval, historical objects and facts clustering, visualization navigation, association analysis, and chronicle facts reconstruction etc.

KEYWORDS

Knowledge Organization, Knowledge Representation, History of Contemporary China, Semantic Mining

1. Introduction

China has a huge volume of historical resources on its contemporary history. With the development of digitization and network technologies, the information resources on the history of contemporary China are increasing in even higher speed, however since most of those information are textual resources that not in a structured manner, many rich semantics and lots of historical knowledge on Contemporary China are hidden in those textual resources and cannot be represented, discovered and utilized easily. It became an urgent problem to mine the implicit knowledge scattered in a large number of historical resources and to represent and organize those knowledge in a fined-grained manner so that it can be used in historical research and education.

Currently, many researcher use a term named "rich semantics" to refer to the implicit knowledge scattered in text resources. In addition to historical resources, many documents

bear rich semantics such as facts, experiences, opinions or other information. Those rich semantics could support a broad range of applications if they can be extracted from text resources automatically or semi-automatically. For example, in medical field, the researchers try to extract the rich semantics from medical texts, to process, represent and store those extracted rich semantics for further analysis (Kerstin Denecke 2016).

In Chinese history field, researchers also try to use semantics technologies to extract, represent and organize rich semantic in document for knowledge discovery. The most commonly used way to represent and organize historical knowledge is ontology. Several ontologies now have been developed to describe and organize historical knowledge, such as "Kuomintang-Communist Cooperation" Historical Ontology (Dong et al. 2006), "Northeast Anti-Japanese Struggles" Historical Ontology (Wu 2012), "Zizhi Tongjian" Historical Ontology (Peng and Song 2010), and "Three kingdoms" Ontology (Liao 2011). Some researchers designed semantic platforms to represent and process the historical knowledge. For example, Prof. Dong and his team constructed the knowledge base based on semantic data from the Chinese Twenty-Four Histories, built the Basic Historical Analysis Platform to discover implicit knowledge in historical records (Dong et al. 2014).

Other works in knowledge organization in history outside China also give us a lot of inspiration. For example, Hyvönen built historical event ontology on Finnish history and developed the semantic portal "CultureSampo", a portal for Finnish Culture on the Semantic Web 2.0 (Hyvönen et al. 2007). Corda proposes a logical model of event ontology for exploring association in history (Corda et al. 2011). Ide and Woolner outlined a model for historical ontologies which is temporally contextualized and could be used to represent relationship of entities in different periods of time (Ide and Woolner 2007).

Based on those related works, this paper proposes a method, which is called "Mining down, Organizing up", to organize fined-grained historical knowledge on Contemporary China. Based on this method, we developed a system to support user to explore historical knowledge by providing functions such as semantic retrieval, historical objects and facts clustering, visualization navigation, association analysis, and chronicle facts reconstruction etc.

2. Framework of "Mining down, Organizing up"

The aim of the project "Knowledge web of the history of the People's Republic of China" is to popularize historical knowledge on contemporary China and promote historical education. One challenge of the project we encounter is to find a way to use semantic technologies to help history experts refine important historical knowledge on contemporary China automatically or semi-automatically from historical resources, such as reference books "*Dictionary of the history of the Chinese Communist Party*", "*Encyclopedia of the National History of the People's Republic of China*", "*Conspectus of Chinese Modern History*", "*Chronicle of the People's Republic of China*" and so on. The other challenge of the project we face is to find a way to organize and represent the historical knowledge so that the users can discover more interesting historical knowledge when they explore the historical knowledge base.

To solve those problems, this paper proposes a method, which is called "Mining down, Organizing up", to semantically represent and organize historical knowledge on contemporary

China hidden in historical encyclopedia text. Based on the proposed historical ontology of contemporary Chinese, this method extracts knowledge objects and facts from the unstructured historical text items by utilizing text mining technologies, represents the historical knowledge in semantically enriched way, and interlinks the related historical knowledge objects and facts to form a historical knowledge network of the contemporary China. By mining the historical facts and the historical knowledge network, the authors get more valuable patterns from the historical knowledge which could be used to form the new organization scheme to reorganize the historical knowledge in a more vivid way.

The framework is shown in figure 1.

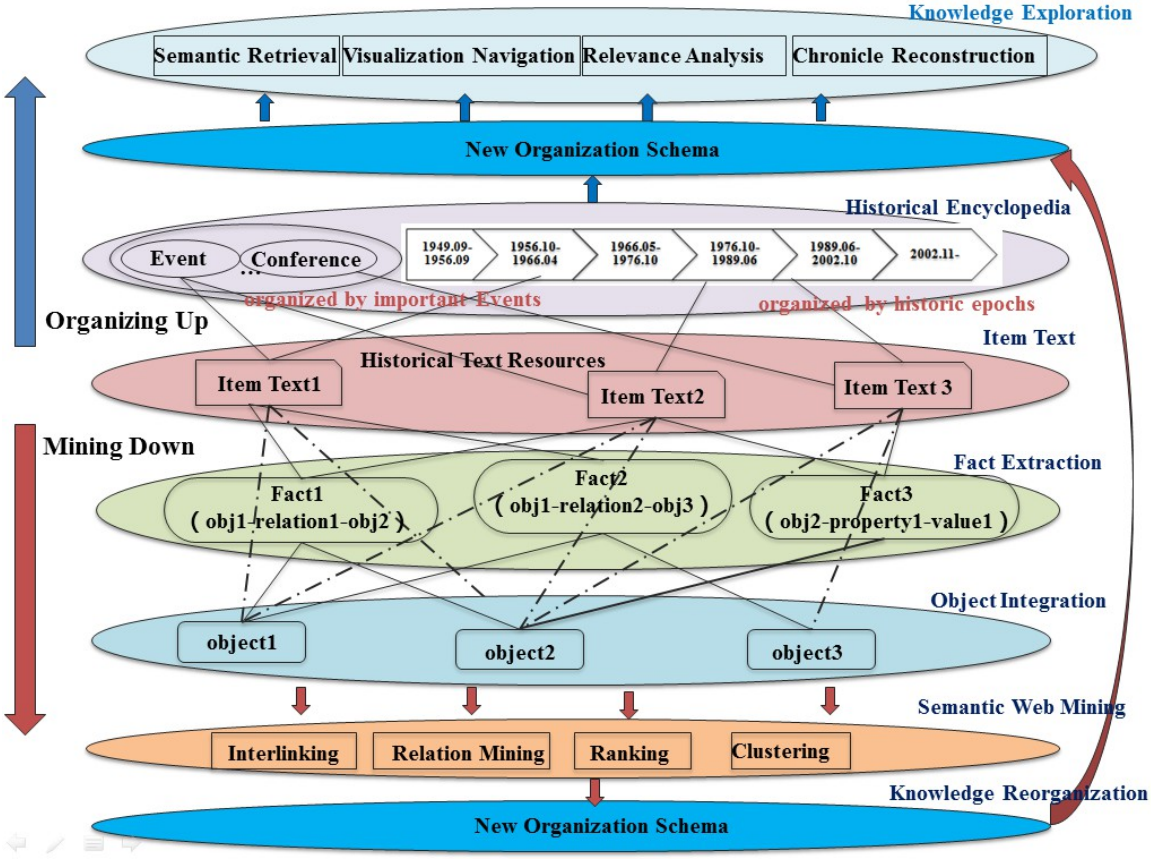

Figure 1 Framework of "Mining Down, Organizing Up"

In detail, "Mining Down" is a process of deconstruction, which transforms knowledge in historical text resources into a series of historical knowledge objects, facts and text items which form the historical knowledge network. By using text mining technology, this process automatically extracts knowledge objects from unstructured item texts, annotates important sentences in the text, and then performs relation extraction to extracts facts from those sentences. All the extracted historical knowledge objects and facts are confirmed or corrected by history experts. So we turn the unstructured item texts in to many structural facts like "object1- relation- object2" or "object1- property- value".

Facts about one knowledge object may be extracted in the same or different text items. Different objects can also be associated directly or indirectly by the relationship (or facts) between those objects. More over a text item can also be interlinked to others according to the object or fact in the text. In this way, the historical knowledge network can be constructed by

the association of knowledge objects, facts and text items. Based on semantic web mining techniques such as ranking, clustering, interlinking, relation mining, pathway analysis, more valuable patterns from the historical knowledge can be achieved and the fined-grained historical knowledge on Contemporary China can be reorganized.

"Organizing Up" is a process of construction, which utilizes the historical knowledge network constructed by the “Mining Down” process and the new organization schema derived from semantic web mining to reorganize the historical knowledge in a more vivid way. Currently, historical text items on the contemporary China are simply organized by historic epochs (such as “1949.09-1956.09” period, “1956.10-1966.04” period, etc.) and organized by important events (such as political event, political convention, foreign affairs, etc.), knowledge hidden in text still cannot be effectively discovered. Based on the constructed historical knowledge network, "Organizing Up" reorganizes historical knowledge by using the patterns and the new organization scheme derived from “Mining Down" process, develop a system to help user explore the historical knowledge through semantic retrieval, visualization navigation, relevance analysis and chronicle facts reconstruction.

3. Methods

There are several key problems need to be resolved in the process of implementation of the proposed method "Mining Down, Organizing Up", which include how to develop basis ontology to describe the objects and relationship in the historical knowledge on contemporary China, how to get core knowledge objects that could be used to guide the objects and facts extraction, how to extract facts about knowledge objects from text resources and how to perform semantic mining to form new organization schema. To solve these problems, this paper proposes the specific methods as follow.

3.1 Ontology definition

There are lots of rich semantics hidden in historical resources on contemporary China. It is necessary to make clear what kinds of rich semantics should be extracted and disclosed to user, namely the knowledge organization model that should be firstly confirmed. Therefore, we build contemporary Chinese history ontology to organize the extracted objects and facts. In contemporary Chinese history ontology, we define the types of core knowledge objects, properties of knowledge objects and relations between objects. Based on concept schema of the ontology, we can outline the knowledge framework on history of contemporary China.

By Referring ontology construction methods such as Skeletal Methodology (Uschold and King 1995) and Seven Steps (Noy and McGuinness 2015), we propose the concept schema of contemporary Chinese history ontology with the help of history experts. By analyzing the text resources, we found that the track of history is mainly consisted of important historical events, conferences, people, etc. Therefore, we firstly define 15 classes in the historical ontology, such as event, conference, person, institution, document, concept, and so on (see figure 2). Secondly, according to the description of these classes, define 20 datatype properties and 76 object properties to model the properties and relationships of knowledge objects. For example, in figure 3, to represent details of historical events, the ontology defines datatype properties: label, alternative label, literal description, and object properties: parent event, subevent, related people, related institution, related event, occurrence

time, occurrence place, etc. Thirdly, define property restrictions, for example parent event property and subevent property are inverse and transitive, label property and nationality property are functional. The detail of historical ontology can be seen in the paper (Sun 2014).

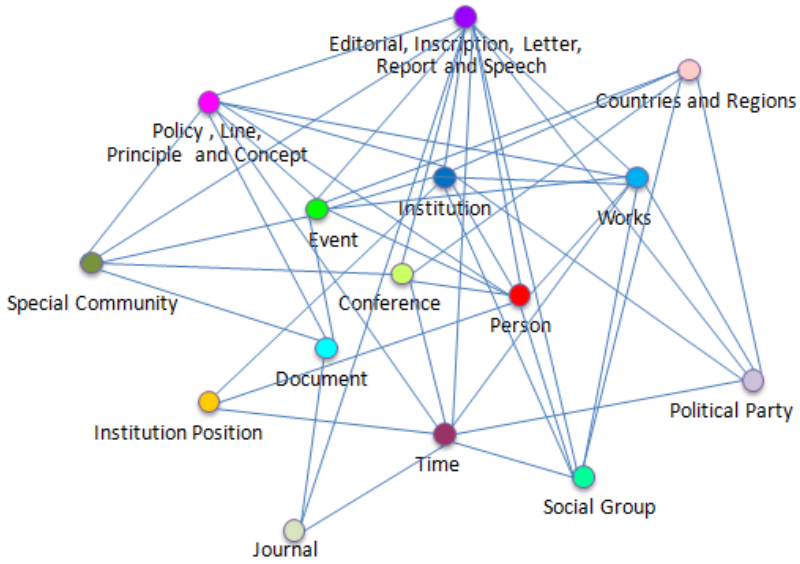

Fig 2 Core Classes in Ontology

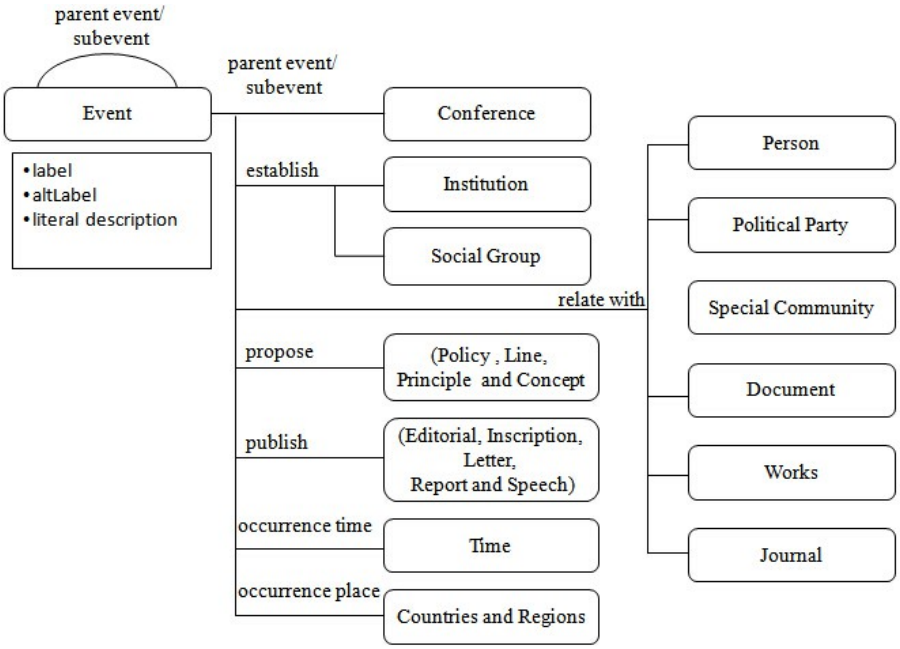

Fig 3 Datatype properties and object properties of Event Class

3.2 Core knowledge objects identification

To populate this ontology, we first extract the metadata of historical information resources and aggregate titles of typed text items which describe an event, a conference, a person, or a document, etc. Moreover, we integrate existing subject headings including person names, institution names, political parties and geography names. Most importantly, we obtain core historical events, conferences and their hierarchies and associations since the founding of the People's Republic of China which has been manually identified and normalized by history experts. These data are core knowledge objects for further representation and organization.

Normalized knowledge objects include 1,685 events, 761 conferences, 3,508 persons, 2,621 institutions, 155 social groups, 107 special communities, and 1,861 hierarchy relations between events or conferences, etc. These knowledge objects are respectively populated into ontology as individuals with URI, standard label, and alternative label. Their relationships are represented by RDF triple statements. In next step, we will use these objects as corpus for semantic mining.

3.3 Fact extraction

"Mining down" method is applied on historical text items to find out the relevant facts of above core objects and to populate the properties and relations of corresponding ontology individuals. In this way, the knowledge hidden in text will be revealed and can be explicit expressed and calculated. Specifically, we use text mining technology to develop automatic processing on text items which helping history expert establish semantic associations between knowledge objects.

(1) Extract knowledge objects

With the aid of knowledge object names dictionary, we execute semantic annotation processing by identifying normal name or alternative names of knowledge object whether appeared in text items. In addition, we develop a named entity recognition tool to discover new knowledge objects such as time, person, institution, conference, etc. and then suggest them to history experts.

(2) Extract facts of knowledge objects

Furthermore, we detect relevant facts of knowledge objects and refine sentences of facts to history experts using relation extraction technology. For example, a text item with title "The Third Plenary Session of Eleventh Central Committee of the Communist Party of China" in "*Encyclopedia of the National History of the People's Republic of China*" describes the content of "the Third Plenary Session of Eleventh Central Committee". The sentence "The third plenary session of eleventh central committee, which held in Beijing on December 18 to 22, 1978, has profound significance in the history of communist party of China since the establishment." implies some facts: the holding time of "The third plenary session of eleventh central committee" is "December 18-22, 1978", the location is "Beijing", etc. According to datatype and object properties defined in the historical ontology, we collect many predicate verbs, such as the "hold", "convene" "take place", etc, and create extraction rules of "conference - time", "conference - location", etc. By using syntactic analysis and relation extraction, the facts of knowledge objects can be extracted from text items.

While automatically processing can be used to find out potential knowledge, due to the complexity of natural language, the accuracy of text mining methods still need to be improved and the results cannot be directly added into the historical ontology. All these need be identified, complemented and revised by experts based on their own domain knowledge.

3.4 Knowledge network construction

After the above steps, the knowledge network, consisting of three layers: text item layer, fact layer and knowledge object layer, can be built. For example, figure 4 illustrates the process of knowledge network construction. The text item "The Third Plenary Session of Eleventh Central Committee of the Communist Party of China" in "*Dictionary of the history*

of the Chinese Communist Party" reveals the facts about holding time, location, present members, and related event of the third plenary session of the communist party of China. The text items with same title in "*Encyclopedia of the National History of the People's Republic of China*" and "*Chronicle of the People's Republic of China*" not only include the above facts but also reveal related concepts of "Emancipate the Mind" and "Seek Truth from the Facts". In the text item "The great historical turning point" of "*Conspectus of Chinese Modern History*", the occurrence time, place, related conference, and related event are displayed, in addition with the facts including related persons and conference of "The 11th National Congress of the Communist Party of China", related persons of "The movement to criticize the 'Gang of four'", and so on. Thus, by the method combined of text mining technology and domain knowledge of history experts, the internal knowledge are discovered from the text, at the same time a complex network of historical knowledge on Contemporary China is constructed.

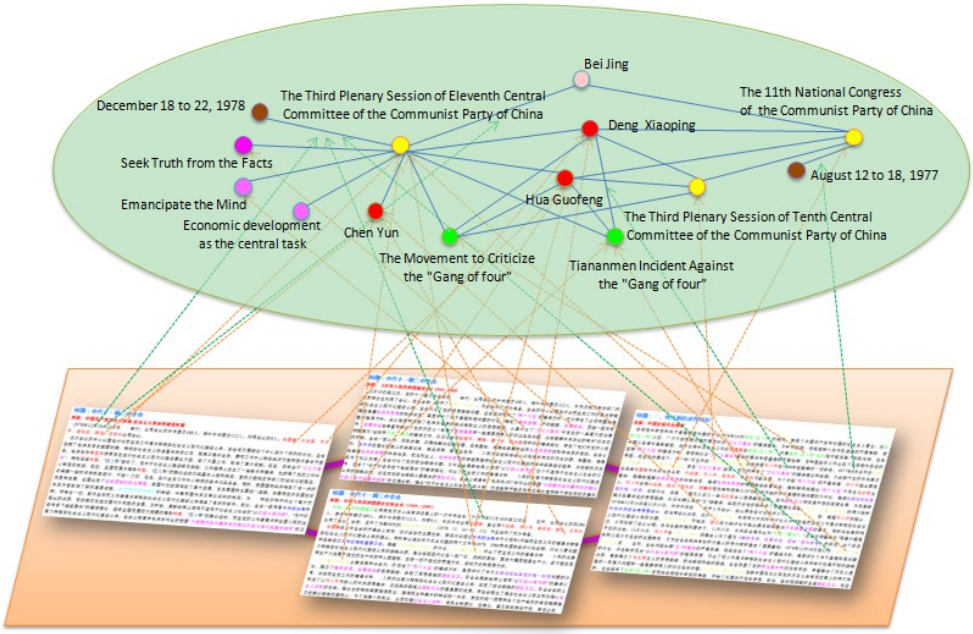

Fig 4 Construction of knowledge network on "The Third Plenary Session of Eleventh Central Committee of the Chinese Communist Party"

3.5 Multidimensional organization

Based on the knowledge network, historical knowledge can be organized in higher level according the relations such as time, subclass, hierarchy, and statistics.

(1) Organization in a time line

Time dimension is the most direct way to show historical development process. For instance, text items from different books can be organized in historical period, knowledge objects and their facts can also be ordered according to time class in ontology. In addition, the facts in the same historical period can be organized together, such as occurrence events, held conference, proposed policy, founding institutions, published works or documents, presented speeches, and so on.

(2) Text item organization based on knowledge objects

The knowledge objects and facts extracted from text items provide a basis for deeper

organization of text items. The same fact from different source items not only verifies its accuracy, but also reflects close relationship between these text items. The text items about the same knowledge object or fact can be organized together for historical biography, institutional evolution, historical data compilation, book writing, etc, which providing references for the history research of contemporary China.

(3) Semantic organization in fact/object dimension

The contemporary Chinese history ontology is used to effective organize historical knowledge and to provide specific semantic representation on historical knowledge objects and facts, moreover it facilitates knowledge exploration including retrieval, association, clustering, reorganization, etc. On the one hand, it supports fine-grained knowledge retrieval, directly to knowledge rather than text resources, as structured query on facts of ontology can be implemented by SPARQL language. On the other hand, the same type of knowledge objects can be gathered by their semantic association, the facts of one knowledge object can be used to build a network describing this object, flexibly develop knowledge integration for various applications.

As mentioned above, "Organizing up" method implements multi-dimensional display of the historical knowledge in higher level. Meanwhile, based on knowledge objects and facts, these text items can also be associated with external resources such as history books, literatures, materials, web pages or databases, and can be used to develop extension applications on historic knowledge of contemporary China.

4 Semantic Applications

We have applied the above method to develop a service platform that provides applications including semantic retrieval, historical objects and facts clustering, visualization navigation, association analysis, and chronicle facts reconstruction etc.

4.1 Semantic retrieval

Different from keyword-based search, we implemented semantic retrieval based on the historic knowledge network. When user submits a search, the system returns the knowledge object which preferred label or alternative label matching the query word. If no matching, it suggests some similar knowledge objects or historical text items. In figure 5, the result of a query entry with "Land Reform Movement" is displayed in a network which represents the object of "Land Reform Movement" linked to its related conference, event, document, person, organization, etc. The historical objects and facts are clustered in the network, and user can directly obtain knowledge which was hidden in text in the past. In addition, if they want, user can also further browse the source texts about this object.

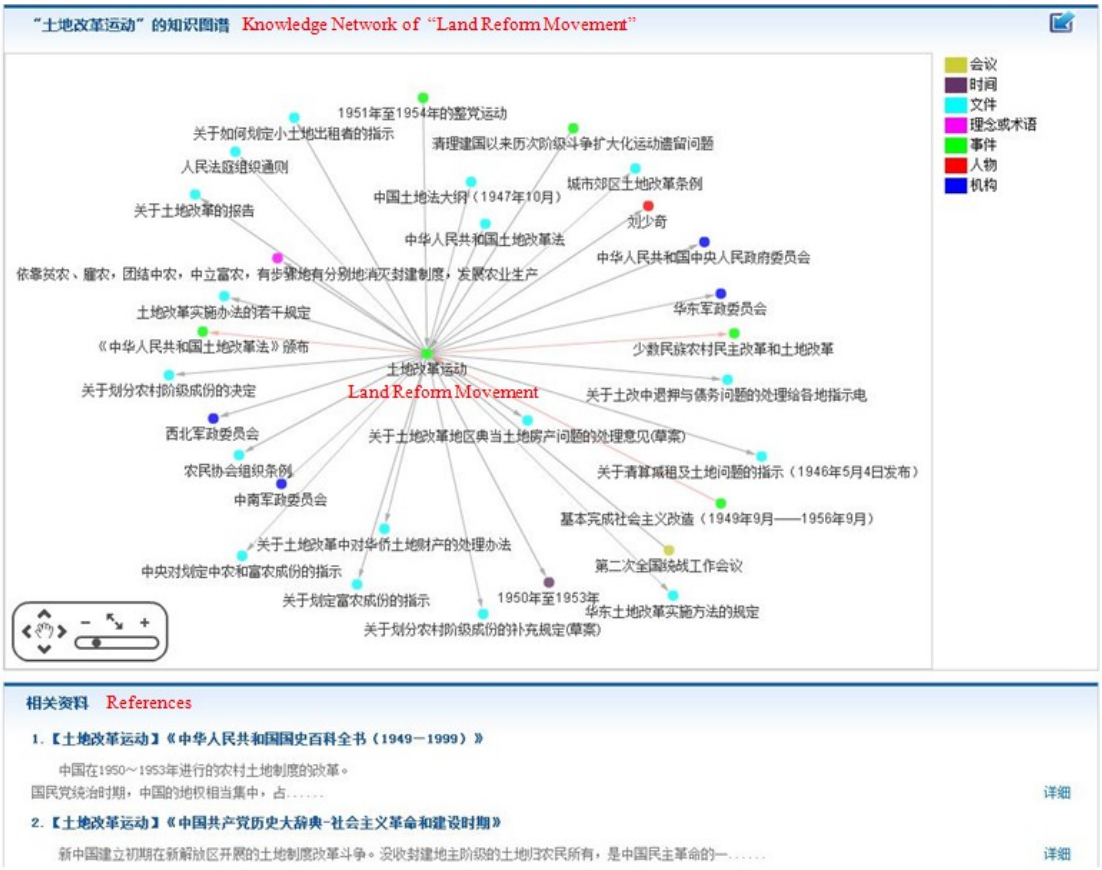

Fig 5 Example of semantic retrieval result

The platform also provides query answering module based on ontology. If user submits a question "Who proposed 'All the reactionaries are the Papertiger'", it will return the answer by a direct edge labeled "proposer" pointed from the node represent the term "All the reactionaries are the Papertiger" to another person node "Mao Zedong" in figure 5.

Fig 6 Example of query answering

4.2 Visualization navigation

Furthermore, we implemented knowledge network visualization, in which nodes represent knowledge objects and edges represent semantic relations (see figure 7). Users can intuitively get wanted historical knowledge without needing to read the text information. The network can also be used as visualization navigation for more knowledge by click on nodes. This improves the efficiency of knowledge acquisition.

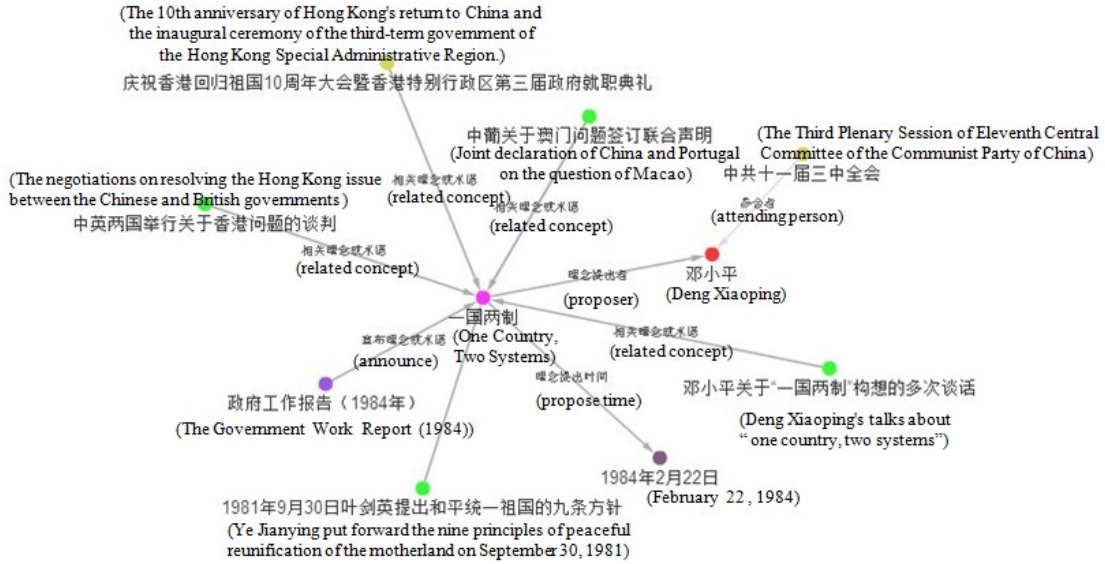

Fig 7 Fragment of visualization navigation

Figure 7 is a fragment of visualization navigation. When browsing the knowledge object "The Third Plenary Session of Eleventh Central Committee of the Communist Party of China", we can see its attended person "Deng Xiaoping". If we click on the node of "Deng Xiaoping", it will display the facts about "Deng Xiaoping", for instance, "Deng Xiaoping" proposed the concept "one country, two systems". If we continue to click on the "one country, two systems", it will show that it is proposed on "February 22, 1984". If we right-click on the edge between "one country, two systems" and "February 22, 1984" with mouse, we can browse its source text item and get its context “On February 22, 1984, when meeting guests of the United States, Deng Xiaoping explicitly proposed the concept ‘one China, two systems’. In the same year on May 15, the government work report passed at the second session of the sixth National People's Congress determined 'one country, two systems' to be basic principle of national reunification.”. In the same way, if we click on “The Government Work Report (1984)”, its details will be displayed in the network. So the knowledge network navigation and browse have been implemented.

4.3 Relevance analysis

Based on the knowledge network, relevance analysis can be used to discover potential knowledge between knowledge objects by graph traversal algorithm. For example, if search the association between “Deng Xiaoping” and “The Third Plenary Session of Eleventh Central Committee of the Communist Party of China” with path length not more than 3, the result knowledge network will be shown in figure 8. It displays their related conferences, documents, events, persons, institutions, etc., and the links between these nodes. There are not only a direct relationship between them, “Deng Xiaoping->attended->The Third Plenary

Session of Eleventh Central Committee of the Communist Party of China", but also more indirect links by using relevance analysis. In this way, relevance analysis base on knowledge network can be applied to discover the potential relationship between knowledge objects and to explore much further for more history knowledge.

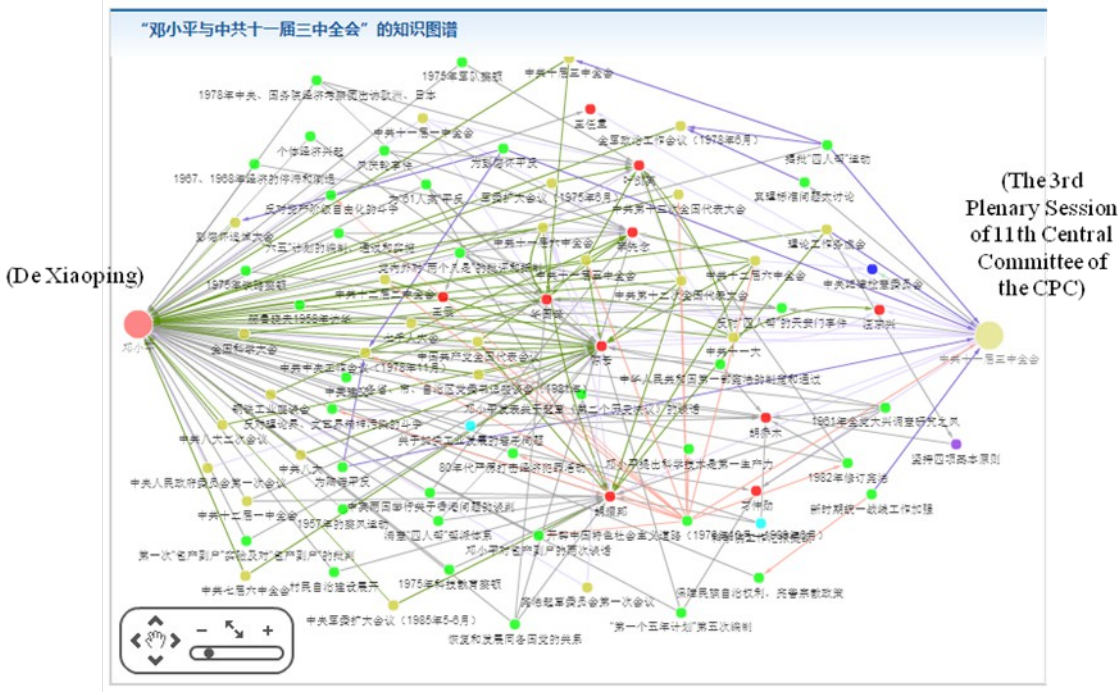

Fig 8 Example of relevance analysis

4.4 Chronicle facts reconstruction

A chronicle is a historical account of facts and events ranged in a time line. In this study, the time class in the historical ontology can be used to represent facts and events accurate to "month" or "day", for instances, the establishment date of political parties, institutions and social groups, the occurrence time of events and conference, the proposed time of principles and concepts, etc. All these information can be used for memorabilia, chronicle or something else. Figure 9 shows the historical activities of Chairman Mao in 1949. These data are computed by indirect relations between the person class and the time class. For example, it is explored that "Mao Zedong" participated the "The first plenary session of the Central People's Government Committee", this conference was held in October 1, 1949. Similarly, all the historical activities can be generated.

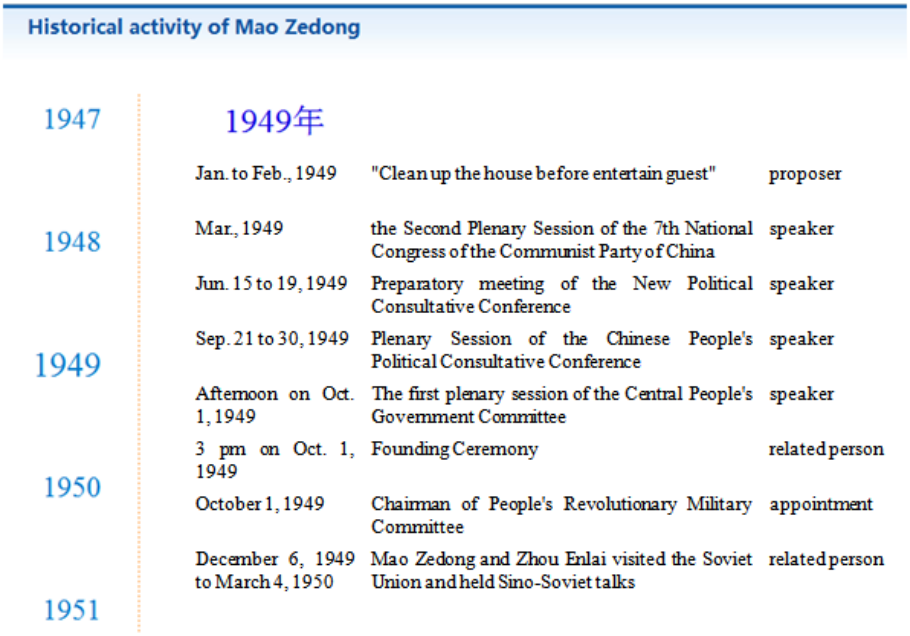

Fig 9 Example of historical activity

5 Conclusions

To implement the effective organization and utilization history knowledge on contemporary China, this paper proposes the method "Mining down, Organizing up". It uses the contemporary Chinese historical ontology for semantic organization and knowledge discovery, and applies text mining technology to extract important knowledge objects and facts from history text items for forming a knowledge network. Based on the knowledge network, this method is applied to develop applications such as semantic retrieval, visualization navigation, association analysis, and chronicle facts reconstruction etc. Studies show that the “Mining down, Organizing up” method can implement fine-grained representation of historical knowledge of the contemporary China and innovative application of knowledge organization based on historical knowledge objects; it can be used as a kind of new organization and exploration method applied in other domains. This study also has some limitations: 1) the accuracy of recognizing knowledge objects and relevant facts from text should be improved, especially the identification of relevant national history facts, it will further reduce the workload of domain experts; 2) The association calculation method of historical knowledge network is simple and has not fully applied current semantic similarity calculation and graph mining methods. These are key problems that need to be solved in our future studies.

Acknowledgments

This article is supported by the project “knowledge web of history of the People’s Republic of China” funded by the Chinese Academy of Social Sciences (Grant No. H1417) and the project “Research on semantic mining of academic resources based on linked data” funded by National Social Science Foundation of China (Grant No. 15CTQ006).

References

［Conference article］Kerstin Denecke, Yihan Deng, Thierry Declerck (2016). Extraction and Processing of Rich Semantics from Medical Texts， *in Joint Proceedings of the 2th Workshop on Emotions, Modality, Sentiment Analysis and the Semantic Web and the 1st International Workshop on Extraction and Processing of Rich Semantics from Medical Texts,* Heraklion, Greece, May 29, 2016.

［Journal article］Dong H., Yu C. M., Yang N., Chen L., Xu G. H., Zhang J. D., et al. (2006). Research on the ontology-based retrieval model of digital library—history domain ontology building. *Journal of the China Society for Scientific and Technical Information*, 2006, 25(5) , 564-474.

［Journal article］Wu L. J. (2012). Research on knowledge organization of characterized database based on ontology. *Journal of Library Science*, 2012(3) , 41-43.

［Journal article］Peng W. M., & Song J. H. (2010). Research on Zizhi Tongjian historical ontology construction and application. *Journal of Chinese Information Processing*, 2010(2), 33-38.

［Journal article］Liao Z. F. (2011). Research on domain ontology constructing and reasoning of the three kingdoms. Central China Normal University, Wuhan, China.

［Journal article］Dong H., Xu L., Wang F., & Yu S. W. (2014). Study on semantic analysis system(Ⅲ) - implementation of Chinese historical records semantic analysis system. *Journal of the China Society for Scientific and Technical Information*, 2014, 33(2), 204-214.

［Conference article］Hyvönen E., Alm O., & Kuittinen H.(2007). Using an ontology of historical events in semantic portals for cultural heritage. In *Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007)* (pp. 1-2). Springer.

［Conference article］Corda I., Bennett B, & Dimitrova V. (2011). A logical model of an event ontology for exploring connections in historical domains. In Proceeding of the *Workshop on Detection, Representation and Exploitation of Events in Semantic Web (DeRiVE 2011), Workshop in conjunction with 10th International Semantic Web Conference (ISWC 2011)* (pp. 1-10). Bonn, Germany.

［Book］Ide N. & Woolner D. (2007). Historical ontologies. In K. Ahmad, C. Brewster, & M. Stevenson (Eds.), *Words and Intelligence II: Essays in Honor of Yorick Wilks* (pp. 37-152). Springer.

［Conference article］Uschold M., & King M. (1995). Towards a methodology for building ontologies. In *Proceeding of the Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCA I-95* (pp. 1-13). Montreal, Canada.

［Web doucument］Noy N. F., & McGuinness D. L. (2015). Development 101: a guide to creating your first ontology. Available at <http://wenku.baidu.com/view/30fb4b956bec0975f465e2bf. html>.

［Journal article］Sun H., &Lei F. (2014). Research on the contemporary Chinese history ontology building. *Journal of Modern Information*, 2014, 34(2):32-42.